Data Science

資料科學期末報告

臺北市房屋租金預測

Group 1
統計三 徐語瑭｜資管三 郭大呈
統計三 陳沛潔｜資科三 林冠儀
統計三 林承佑｜資科四 潘煜智

執行摘要

專案目標

透過建立模型預測臺北市租屋「總金額」

專案背景

臺北市房租高昂，對學生與上班族都是沉重的負擔，且租屋市場資訊不透明，讓租屋決策變得困難。因此我們希望透過建立臺北市租屋價格預測模型，以預測租賃物件的「總額元」做為目標，透過數據分析提供合理價格參考

資料來源

擷取內政部「不動產交易實價查詢服務網」的資料，取得包括建物標的、建物面積、屋齡、總價格及格局配置等資料共 12,271 筆，時間跨度為 14 年。並額外加入「與最近捷運站的距離」作為欄位，以提升模型的預測準確性

執行流程

資料清理	建立模型	特徵工程	模型評估	建立網站
• 資料清洗 • 類別欄位簡化 • 欄位篩選	• LightGBM • XGBoost	• 資料拆分 • 特徵篩選 • 超參數調整	• 比較模型效能	• 使用 Shiny 建立視覺化網站

衡量指標

RMSE、MAPE、MEAPE

專案成果

相較於 Null Model，LightGBM 的 RMSE 下降 51.7%，XGBoost 的 MAPE 及 MEAPE 分別下降 57.2% 與 52.6%

資料集簡介

資料來源

透過爬蟲，擷取內政部「不動產交易實價查詢服務網」的租屋資料，共 12,271 筆，區間為 2012~2025 年

原始資料欄位　共36個欄位

	欄位名稱	欄位說明
位置與編號	編號	租賃資料編號，唯一識別每一筆資料
	土地位置建物門牌	建物地址
	鄉鎮市區	建物鄉鎮市區
交易相關	交易標地	交易涉及的房屋、土地、車位或租賃房屋
	總額元	總價格，為最終預測目標值
	租賃筆棟數	移轉登記實際交易之筆棟數及車位數，包含土地、建物、車位。 資料格式 ：土地0建物1車位0
	單價元平方公尺	每平方公尺單價
	移轉編號	同一筆交易中多項資產的流水編號
車位	車位類別	說明車位型態（如坡道平面、升降機械等），無車位資料多
	車位總價元	車位總交易金額（單位：新台幣）
	車位面積平方公尺	該交易中車位的總面積，單位為平方公尺
土地	都市土地使用分區	分為住、商、都市，且都市類別的括號內有更詳細分類。
	土地面積平方公尺	土地所有權變更所涉及的土地面積加總，單位為平方公尺
	非都市土地使用編定	皆為空值
	非都市土地使用分區	

	欄位名稱	欄位說明
租賃資訊	出租型態	欄位值為「整棟（戶）出租」、「獨立套房」、「分租雅房」等
	租賃期間	租賃期間，時間單位為天數
	主要用途	建物用途e.g. 住家用、商業用、一般事務所
	租賃年月日	租賃開始民國日期（簽約日） e.g.1031106
	租賃層次 / 總樓層數	租賃層次：樓層e.g.五層、六層 / 總樓層數：建物總樓層數
建物	建物型態	建物的類型，包括「住宅大樓 」、「公寓」、「透天厝」 等
	建物完成年月	建築完成民國日期，e.g. 1140401
	主要建材	建築使用材料，如鋼筋混凝土造、加強磚造等
	建物現況格局-房/廳/衛/隔間	房：房間數量 / 廳：客餐廳數量 / 衛：衛浴數量/ 隔間：是否有隔間
	主建物總面積 / 陽台面積 / 附屬建物面積	主體建物：指不含陽台與附屬建物 附屬建物如車庫、地下室、騎樓等附加空間　面積單位：平方公尺
服務與設備	有無電梯 / 有無管理組織 / 有無管理員 / 有無附傢俱	資料欄位值：有/無
	租賃住宅服務	服務名稱，如社會住宅代管、一般包租等
	附屬設備	附屬設備名稱，資料格式如「冰箱、冷氣」、「有線電視、洗衣機」

資料集簡介

新增捷運資料欄位

為探討附近捷運站對於房屋租金的影響，新增了以下兩個欄位

欄位名稱	說明
捷運站距離(公尺)	建物與最近的捷運站的直線距離，以公尺計算
捷運線	離建物最近的捷運站站點屬於哪一條捷運路線，如文湖線、板南線

資料獲取方式

地址座標轉換

利用 ArcGIS 地理編碼服務將建物的地址轉緩成經緯度座標

捷運站點蒐集

從台北市資料大平台搜集台北捷運各出入口的經緯度座標，並注記該站隸的捷運線

距離比對

將建案的座標與所有捷運站出入口比對，找出距離最近的捷運出入口，並將相關欄位加入原本的資料集

資料前處理

資料清洗

欄位調整 → 欄位值調整 → 資料刪除

1 欄位移除

欄位名稱	移除原因
非都市土地使用分區、非都市土地使用編定	原始資料中欄位皆為空值
附屬設備、租賃期間、建築完成年月	後續將欄位切分或重新計算，因此將原欄位刪除

2 欄位拆分/新增

欄位名稱	說明
交易比棟數_土地 / 建物 / 車位	原始資料格式為「土地0建物1車位0」，將欄位拆分為 「交易比棟數_土地」、「交易比棟數_建物」、「交易比棟數_車位」，欄位值則對應交易項目後面的數字
附屬設備	將欄位切分為「附屬設備_{設備名稱}」的格式，e.g.「附屬設備-電視機」 若房屋有該設備則欄位值為1，無則為0。

資料前處理

資料清洗

欄位調整 | **欄位值調整** | 資料刪除

欄位名稱	說明
建築完成年月 租賃完成年月	將日期格式改為YYYY-MM-DD 之字串，若欄位值解析失敗則刪除該列
租賃期間天數	將起迄日相減轉為租賃天數，若計算失敗則補上為（ "NA"）
租賃層次	原始資料格式為「{數字}層」，調整為數字　e.g. 「三層→ 3」、「四層→ 4」… 若原始資料值為其他值，則依以下規則更改「地下X層→ −X」、「地下層→ 0」、「其他 → "NA"」
主要建材	將缺失值補為 ("NA")
總額元	將欄位值去除逗號並轉為數字，若調整失敗則刪除該列

資料前處理

資料清洗

欄位調整 | **欄位值調整** | 資料刪除

欄位名稱	說明
有無管理組織	將是否（布林值）轉為0/1，缺失值則補 "NA"
有無管理員	
有無附傢俱	
有無電梯	
建物現況格局-隔間	

資料前處理

資料清洗

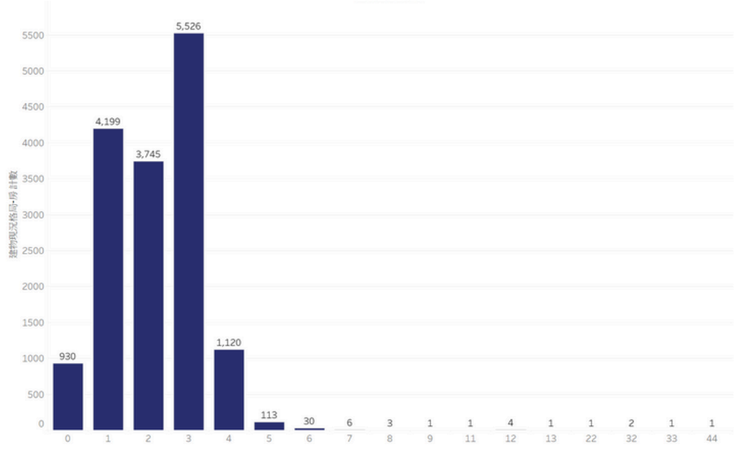
欄位調整 → 欄位值調整 → **資料刪除**

欄位名稱	說明
主要用途	僅留下以下欄位值之資料 「住家用／住宅／集合住宅／多戶住宅／國民住宅／公寓／雙併住宅／農舍／住商用／住工用／宿舍／寄宿／住宿單元」
建物現況格局-房 建物現況格局-廳 建物現況格局-衛	將欄位值大於 100 的列刪除

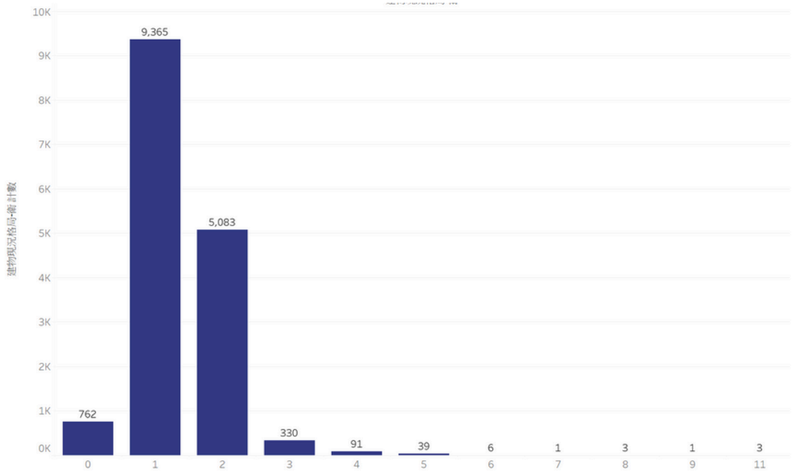
資料 EDA

資料欄位 EDA

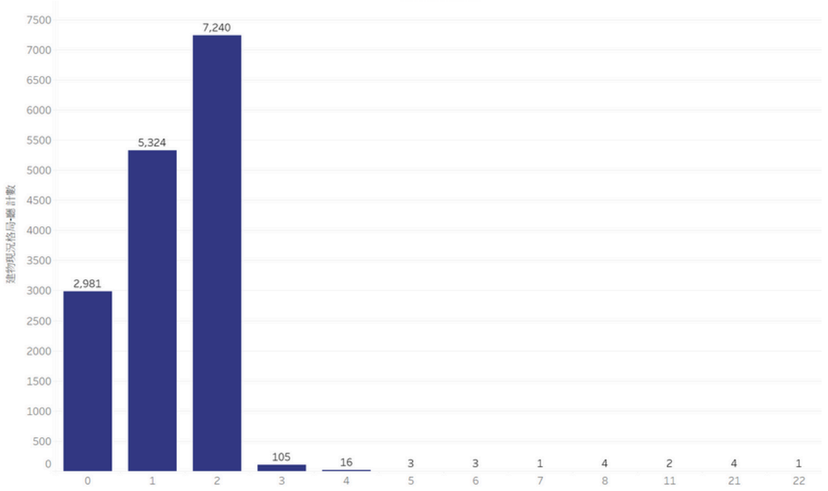
建物現況格局-房



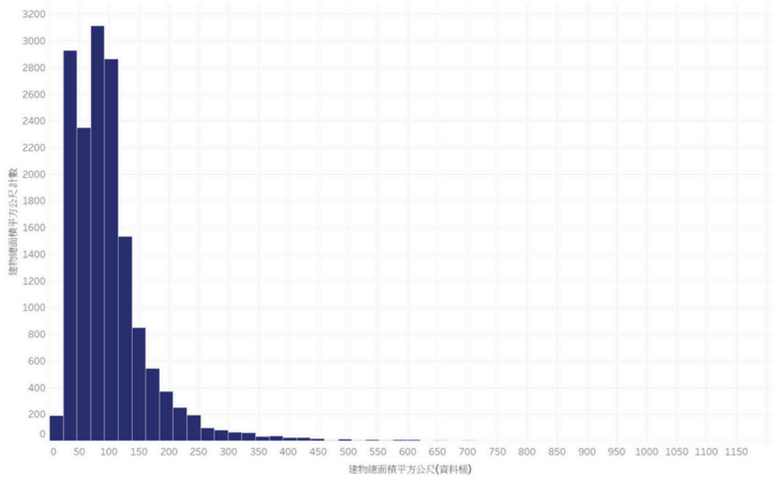
建物現況格局-衛



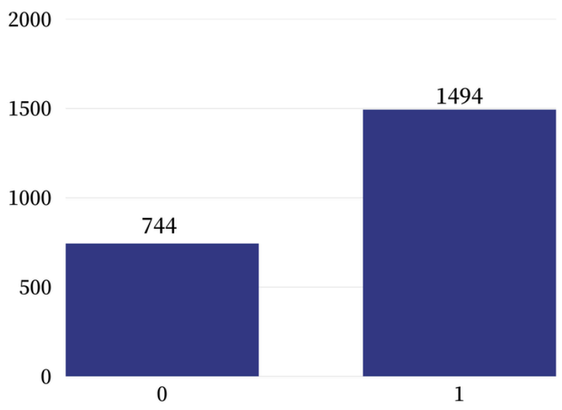
建物現況格局-廳



建物面積



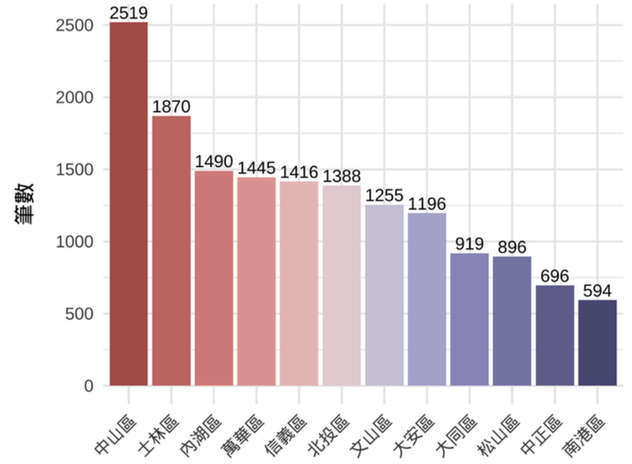
建物現況格局-隔間



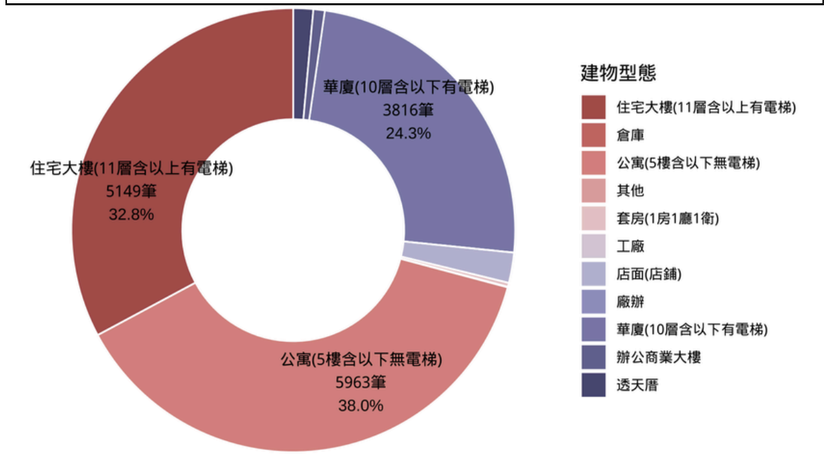
資料 EDA

資料欄位 EDA

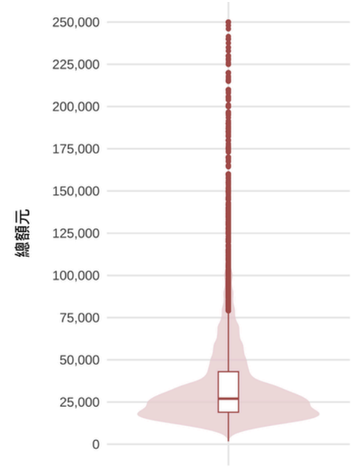
各行政區租屋資料筆數



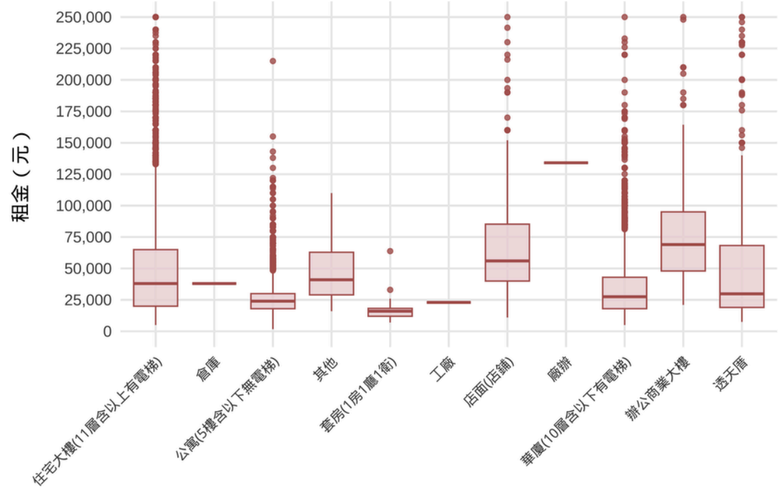
不同建物類型數量分佈



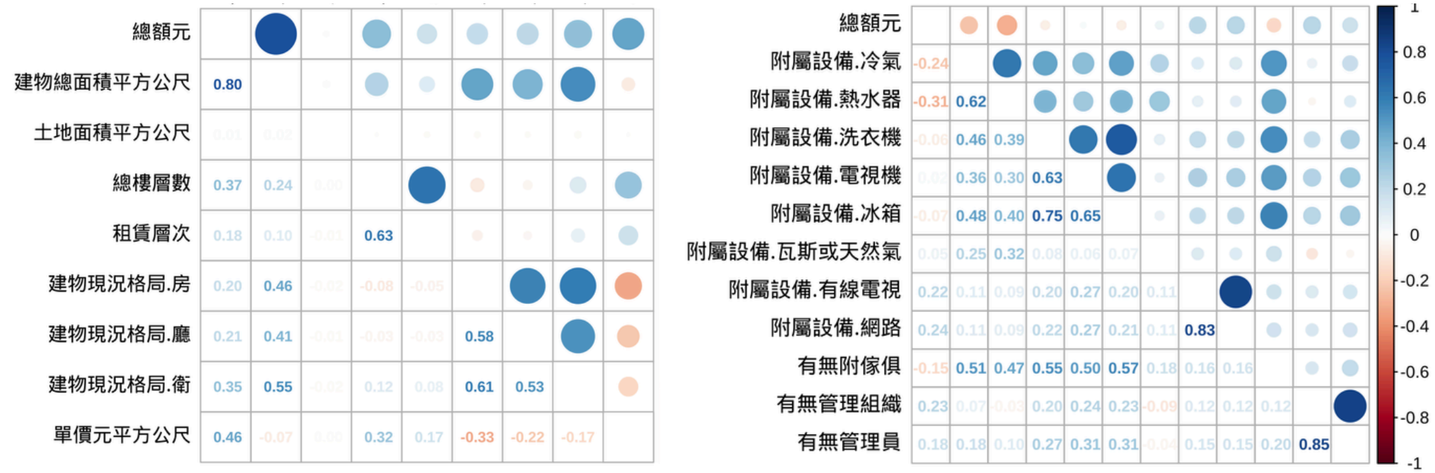
租金分佈箱型圖



不同建物類型租金分佈



相關係數熱力圖



特徵工程

特徵工程

根據資料前處理與 EDA 結果，初步篩選出具有預測能力的特徵變數，作為後續建模的候選欄位。

類別欄位簡化與資料篩選 → 特徵選擇 → 新增特徵

1 類別欄位簡化

原始資料中使用 one-hot encoding 表示捷運路線（如「文湖線」、「淡水信義線」等），我們將其整合為單一「捷運線」類別變數，並將未涵蓋之物件歸為「無捷運」，以簡化欄位結構、降低模型維度並減少共線性問題。

2 資料篩選

為排除不具代表性的短期租賃案例，我們僅保留「租賃天數 ≥ 30」的樣本作為訓練與預測資料，避免因短租價格結構不同而混淆模型。

特徵工程

特徵工程

根據資料前處理與 EDA 結果，初步篩選出具有預測能力的特徵變數，作為後續建模的候選欄位。

類別欄位簡化與資料篩選 → **特徵選擇** → 新增特徵

房屋位置

- 鄉鎮市區
- 土地位置建物門牌

交易相關

- 交易標地
- 租賃年月日
- 單價元平方公尺
- 交易筆棟數__土地
- 交易筆棟數__建物
- 交易筆棟數__車位

租賃資訊

- 出租型態
- 租賃期間天數
- 主要用途
- 租賃層次
- 總樓層數

建物

- 建物完成年月
- 建物總面積
- 建物型態
- 主要建材
- 建物現況格局-房
- 建物現況格局-衛
- 建物現況格局-廳
- 建物現況格局-隔間

車位

- 車位類別
- 車位總價元
- 車位面積平方公尺

租賃住宅服務與設備

- 租賃住宅服務
- 有無電梯
- 有無管理員
- 有無管理組織
- 有無附傢俱
- 附屬設備-冷氣
- 附屬設備-電視機
- 附屬設備-洗衣機
- 附屬設備-冰箱
- 附屬設備-瓦斯或天然氣
- 附屬設備-網路
- 附屬設備-有線電視
- 附屬設備-熱水器

土地

- 都市土地使用分區
- 土地面積平方公尺

特徵工程

特徵工程

根據資料前處理與 EDA 結果，初步篩選出具有預測能力的特徵變數，作為後續建模的候選欄位。

類別欄位簡化與資料篩選 → 特徵選擇 → **新增特徵**

新增「附近建物單價平均價格」欄位

考量到房價會因「地段」而顯著變化，也就是平常說得「這裡的房價一坪多少錢」，因此新增「附近建物單價平均價格」欄位，用以量化「地段」價值。蒐集該筆交易周邊 300 公尺範圍內所有建物的「單價（元／平方公尺）」，對這些單價取平均。

模型選擇與訓練

模型選擇

Method

使用梯度提升樹類模型：適合處理非線性特徵、缺值、類別變數。

LightGBM、XGBoost

Null Model

Baseline 模型：每筆資料的「附近建物單價平均價」乘以「建物總面積」作為預測依據，基準比較。

模型訓練

CV or extra separated data

optimal result of the hyperparameters

1

為避免模型過擬合與資料分布偏斜，依租金進行分箱後再進行分層抽樣 (80%訓練 / 20%測試)。

2

使用 10-fold Cross Validation 確保穩定性與代表性。

模型選擇與訓練

模型選擇

Method

使用梯度提升樹類模型：適合處理非線性特徵、缺值、類別變數。

LightGBM、XGBoost

Null Model

Baseline 模型：每筆資料的「附近建物單價平均價」乘以「建物總面積」作為預測依據，基準比較。

模型訓練

CV or extra separated data → optimal result of the hyperparameters

超參數調整

- 針對 LightGBM 調整以下超參數範圍： num_leaves、learning_rate、 min_data_in_leaf、 feature_fraction、nrounds
- 針對 XGBoost 調整以下超參數範圍： eta、max_depth、subsample、colsample_bytree

特徵篩選

特徵篩選

透過模型判斷特徵貢獻度，挑選對預測最有幫助的特徵。

重要程度判斷

使用 LightGBM 中的 importance 函數，根據每個特徵在決策過程中對降低預測誤差的貢獻程度，給予一個「重要性分數」Gain，這個分數越高，表示該特徵對模型的預測越有幫助。為避免資料洩漏，僅使用訓練資料做計算。

前三大特徵

前三大特徵共貢獻 73.26%

- 建物總面積平方公尺（35.4%）
- 附近建物單價均價（30.0%）
- 屋齡（7.7%）

前 20 個特徵

模型的前 20 個特徵（包括捷運距離）共貢獻了約 99% 的 Gain，這些特徵是模型預測的主要依據。

捷運站距離

「捷運站距離(公尺)」特徵，排在模型中排名第八，Gain 約為 2%。證明了捷運距離對租金預測具有實質貢獻。

顯著性檢驗

透過顯著性檢定來確認是否移除某些特徵後，模型效能真的有統計上的改善。

檢定比較

為找到可以顯著提升效能的模型，針對這 20 種模型去跑顯著檢定

- new model: 移除 n 個最小 Gain 的 feature 的 LightGBM 模型
- base model: 使用全部 feature 的 LightGBM 模型

定義

RMSE 差值：$\triangle = RMSE_{base} - RMSE_{new}$

重抽樣

對測試集做重抽樣（bootstrap）設置 B = 30, 000 次有放回抽樣，每次重算 Δb，取其 2.5%/97.5% 分位作 95% 置信區間，並計算：$2min\{P(\triangle_b \leq 0), P(\triangle_b \geq 0)\}$

結果

p-value 全部大於顯著水準（如 0.05），代表沒有一組 feature 被移除後會讓模型在測試集的 RMSE 和其他指標有統計上的顯著改善。

模型評估指標

評估指標

指標	公式	定義	說明	特色
RMSE Root Mean Squared Error	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$	衡量預測值與實際值之間的平方誤差的平均後再開根號	• 預測成交價與實際值的平均差額 • RMSE = 10000 → 平均誤差為 10,000 元	• 優點：單位直觀 • 缺點：對極值敏感、對資料量級依賴性大
MAPE Mean Absolute Percentage Error	$\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\hat{y}_i - y_i}{y_i}\right\|$	預測誤差與實際值的比例的平均值	• 預測成交價的誤差在預測值的平均佔比 • MAPE = 15% → 平均誤差為實際值的 15%	• 優點：對極值較不敏感、對資料量級依賴性小 • 缺點: 對極端值略敏感
MEAPE Median Absolute Percentage Error	$Median\left(\left\|\frac{\hat{y}_i - y_i}{y_i}\right\|\right)$	樣本的「百分比誤差」的中位數	• 預測成交價的誤差百分比的中位數 • MAPE = 15% → 50% 預測誤差低於 10%	• 優點：對極值最不敏感

指標選擇

- 為進行模型效能評估，模型採用 RMSE 作為主要指標，用以指導交叉驗證與模型選擇。RMSE 可衡量預測值與實際值之間的平均差異，忠實反映整體誤差程度，且單位與預測目標一致（元），利於模型優化與早停條件設計。
- 然而，RMSE 對極端值敏感，且難以直觀理解預測比例誤差。因此，我們補充使用 MAPE、MEAPE 兩個無單位的比例型指標作為模型效能的輔助說明，以全面掌握模型在不同場景與資料分布下的品質

模型成果

模型成效

效能最差
效能最佳

	Null Models	XGBoost	LightGBM
RMSE	19006.05	11970.96	9172.41
MAPE	0.3572	0.1530	0.1538
MEAPE	0.2405	0.1140	0.1191

結論

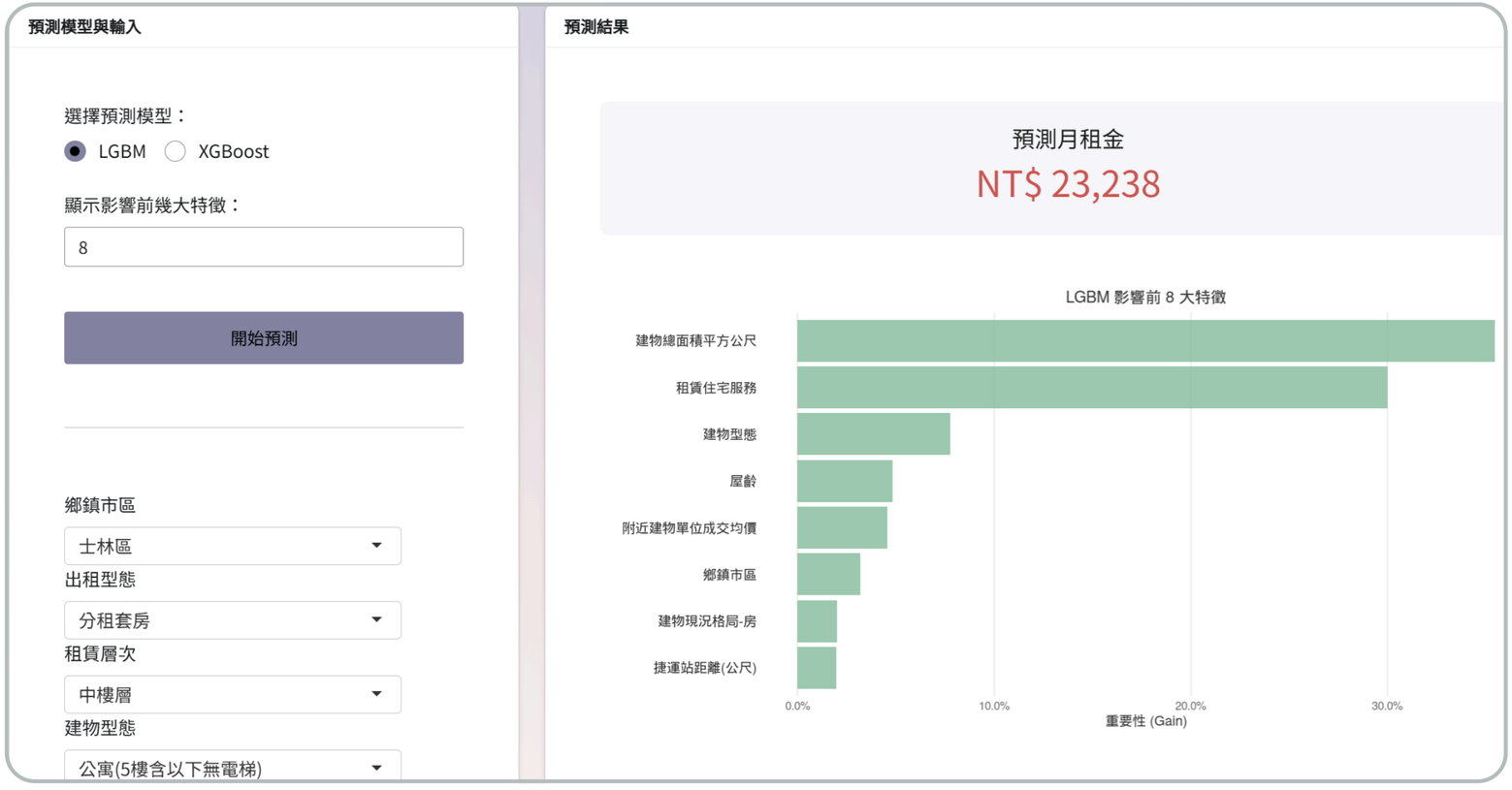
無論採用哪種評估指標，LightGBM 與 XGBoost 相較於 Null model 都有顯著提升，平均預測誤差降低近 10,000 元（約 20%）。其中，LightGBM 在 RMSE 上表現最佳，意味著它最擅長控制整體平方誤差；XGBoost 則在 MAPE 與 MEAPE 上稍微領先。總體而言，若重視絕對誤差的最小化，可優先選用 LightGBM；若更在意相對誤差或極端值的準確性，則 XGBoost 是較佳選擇。

網站 Demo

可互動網站

透過 Shiny 將專案內容以網頁方式呈現，讓使用者可以查看資料說明、數據 EDA、建模架構，並且實際進行操作，輸入不同參數並預測房價

網站連結：https://lgyeee.shinyapps.io/mrt_rental_prediction/



THANKS